GENOMICS 34, 323–327 (1996)
ARTICLE NO. 0293

Characterization and Fine Localization of Two New Genes in Xq28 Using the Genomic Sequence/EST Database Screening Approach

SARA FARANDA,[1] ANNALISA FRATTINI, ILEANA ZUCCHI, CRISTINA PATROSSO, LUCIANO MILANESI, CRISTINA MONTAGNA, AND PAOLO VEZZONI

Istituto di Tecnologie Biomediche Avanzate, Consiglio Nazionale delle Ricerche, via Ampère 56, 20131, Milan, Italy

Received November 15, 1995; accepted March 12, 1996

Two new genes were identified and mapped by searching the EST databases with genomic sequences obtained from putative CpG islands of the rodent–human hybrid X3000. Previous mapping of these CpG islands in the proximity of the host cell factor (HCFC1) and GdX genes automatically localized these two new genes to Xq28 in the interval between the L1 cell adhesion molecule (L1CAM) and the glucose-6-phosphate dehydrogenase (G6PD) loci. Both genes are relatively short, contain an ORF of 261 and 105 amino acids, respectively, and are ubiquitously expressed. Combining sequencing of selected CpG islands, derived from hybrids containing small portions of the human genome, with an EST database search is an easy method of identifying and mapping new genes to specific regions of the genome.

INTRODUCTION

The Xq28 chromosomal region is one of the portions of the human genome richest in gene content, particularly in genes responsible for human hereditary diseases (Nelson *et al.*, 1995). However, the distribution of detected transcripts and/or CpG islands, which are known to be associated mainly with housekeeping genes, is not homogeneous along this chromosomal band. In some portions of the Xq28 band, genes appear to be extremely densely packed (Pilia *et al.*, 1993; Palmieri *et al.*, 1994). In particular, the region between the L1 cell adhesion molecule (L1CAM) and factor VIII (F8) genes has been extensively investigated by many groups, and two contigs, one containing the L1CAM and Host cell factor (HCFC1) genes and the other containing the color vision (CV) and glucose-6-phosphate-dehydrogenase (G6PD) genes, have been assembled (Nelson *et al.*, 1995). Although these two contigs have not yet merged, many transcripts have been detected (Bione *et al.*, 1993; Sedlacek *et al.*, 1993; Tribioli *et al.*, 1994), some of which have been completely sequenced and characterized.

In the past few years, our laboratory has been particularly interested in the region between the L1CAM and G6PD genes, and we have contributed to the identification and mapping of several genes in this chromosomal segment (Frattini *et al.*, 1993, 1994; Patrosso *et al.*, 1994; Faranda *et al.*, 1995). In particular, we mapped the HCFC1 gene to the region located between L1CAM and Methyl-CpG binding protein MECP2 (Frattini *et al.*, 1994; Wilson *et al.*, 1995; Nelson *et al.*, 1995) in close association with the Renin-binding protein (RENBP) gene, which is centromeric with respect to HCFC1 (Faranda *et al.*, 1995). We have now detected a new gene about 2 kb downstream from the HCFC1 gene start site by using an approach that couples genomic sequencing and EST database searches. With the same approach we also identified another EST, which we mapped close to the GdX gene (Bione *et al.*, 1993) in the interval between the CV and the G6PD genes.

MATERIALS AND METHODS

Primers for PCR and RT-PCR and amplification conditions of ITBA2. The primer sequences were as follows (F and R refer to their forward or reverse orientation): ITBA2F, GGAGGCGCTGACGGCGGGGATG; ITBA2R, CCTCCCAGGAAAGTAGCAAC. Thirty cycles of amplification were performed; the denaturation step was at 94°C for 1 min., the annealing step was at 66°C for 1 min., and the polymerization step was at 72°C for 1 min.

Subcloning, DNA sequencing, and sequence analysis. Sequences of genomic regions were compared to EST databases (NCBI–GenBank, performed in September, 1995) with the BlastN algorithm (Altschul *et al.*, 1990), and two sets of clones (ITBA1 and ITBA2) were detected. In the ITBA1 set, two clones that were obtained from Généthon were detected: HSCZXE10 was from an infant brain library and HSBB5D05 was from a skeletal muscle library. In the ITBA2 set, four EST clones, which gave the sequence of a short transcript when assembled, were found as the sequence from both ends merged together. We did not request any clone for ITBA2, but the transcript was obtained by RT-PCR performed with ITBA2F and ITBA2R primer pairs on human liver RNA purchased from Clontech. The sequence performed on both strands of an RT-PCR clone was considered as the true sequence, which was compared to sequences of EST clones.

Sequencing was performed by the dideoxynucleotide chain termination method. PCR amplification products were cloned in TA vector

Sequence data from this article have been deposited with the EMBL/GenBank Data Libraries underAccession Nos. X92475 (ITBA1) and X92896 (ITBA2).

[1] To whom correspondence should be addressed. Telephone: +39 2 70643380. Fax: +39 2 2663030.E-Mail: frattini@itba.mi.cnr.it.

0888-7543/96 $18.00

(Invitrogen) and sequenced using the Sequenase kit (USB). For the sequencing, DNA was prepared with the Wizard Miniprep System (Promega). Exon boundaries were determined by comparing the obtained sequences to the genomic sequence, where available. Computer analysis of nucleotide and amino acid sequences was performed with the Wisconsin Package Version 8.0-UNIX. Homology searches were performed in GenBank (Release 91, 10/95), EMBL (Release 44, 9/95) and PIR Protein (45, 7/95), with FastA.

DNA and RNA extraction and Northern blot analysis. Genomic DNA for Southern blot and PCR amplification was extracted with the phenol–chloroform method. For Northern analysis, filters, purchased from Clontech, were hybridized overnight at 42°C with formamide, with a probe generated by random priming (Feinberg and Vogelstein, 1983). Filters were washed several times in 2× SSC, 0.05% SDS at room temperature for 40 min and in 0.1× SSC, 0.1% SDS at 50°C for 40 min and finally exposed to X-ray film (DuPont NEN).

RESULTS

Cloning and Mapping of ITBA1

The genomic region flanking the HCFC1 gene at the 5′ end has been completely sequenced. This region is extremely rich in G+C and represents a CpG island (Zoppè *et al.*, 1996). Search of databases with this genomic portion only recently detected two ESTs, one from a skeletal muscle library and the other from an infant brain library (subsequently, an additional EST clone, t35674, from a small intestine cDNA library, was found.). Both clones were obtained from Généthon, and their complete sequences were determined (ITBA1). The clones contained a poly(A) tail at their 3′ end, and the length of the determined sequence (1411 bp) is in keeping with the length of the transcript detected with Northern blot (see below). However, no attempt was made to define the exact 5′ end of the transcript.

The sequence of the cDNA is shown in Fig. 1 (top). An ORF of 261 aa is present in the cDNA sequence, starting at nt 285, preceded by a 5′ UT region of 284 nt and followed by 343 nt of 3′ UTR. A polyadenylation site is present at nt 1350–1355. The sequence does not detect any obvious similarity in databases.

Comparison to the promoter region of HCFC1 showed complete identity with the first exon of the ITBA1 gene (first 72 nt of the ITBA1 cDNA). All the other exons lie outside cosmid 430 (Fig. 2), because analysis of the cosmid portion downstream from the first exon did not detect other cDNA fragments. At the same time, this analysis allowed us to establish that ITBA1 maps approximately 2000 bp telomeric from the start site of the HCFC1 gene and that the direction of the transcription of ITBA1 is opposite to that of the HCFC1 gene (Fig. 2). Analysis of the genomic sequence upstream from the start site did not detect either a TATA box or a CAAT box (data not shown), as often happens in housekeeping genes.

A Northern blot with various fetal and adult human tissues was hybridized to ITBA1 cDNA. The expected 1.4-kb transcript was detected in all tissues, suggesting a ubiquitous pattern of ITBA1 gene expression (Fig. 3A).

Cloning and Mapping of ITBA2

The second gene was mapped with a similar approach. We have previously reported the characterization of *bona fide* CpG islands in the Xq24–qter region and the sequencing of some portions of the corresponding genomic clones (Tribioli *et al.*, 1992). At that time, there were very few ESTs in databases, and similarities were detected only with well-characterized genes. We have recently performed an EST database search, and we found that one of these genomic clones matched various ESTs in databases. In particular, the first 177 nt of EST yu55e11 (from Soares' fetal liver–spleen cDNA library) was very similar to the sequence found on the *Eag*I side of a *Eag*I/*Eco*RI genomic segment of 7 kb detecting the GdX gene on the *Eco*RI side (Tribioli *et al.*, 1992), and this allowed us to map the new gene, ITBA2, to the CV–G6PD gene interval, 5 kb from the end of the GdX gene (Fig. 2). By relating the position of the ITBA2 sequence to the known GdX gene orientation of transcription (Bione *et al.*, 1993), we were able to establish that ITBA2 is centromeric to GdX and is transcribed in the same orientation.

By further searching of the EST database, three additional clones (ye56c01, yi95c02, and yt85b05 from Soares' fetal liver–spleen, placenta, and pineal gland cDNA libraries, respectively) were found to contain ITBA2-related sequences. They are shown in Fig. 4, together with the genomic sequence used to probe the database and the sequence obtained by RT-PCR amplification performed on human liver RNA with primers ITBA2F and ITBA2R. In this figure it can be seen that all the sequences pertain to the same gene and not to putative homologues and that the few discrepancies among them can be easily attributed to sequencing errors, whose frequency varies from clone to clone. The RT-PCR sequence was taken as a reference sequence, as it was carefully determined from both strands. The most divergent clone was ye56c01, which accounts for most of the discrepancies shown in Fig. 4, followed by the yi95c02 clone. However, there is little doubt that all the EST clones pertain to the same gene located in the same chromosomal region from which the CpG island used for the database search was derived.

The assembled sequence obtained by RT-PCR and its theoretical translation are reported in Fig. 1 (bottom). The transcript is very short, only 599 bp, but contains a poly(A) tail, suggesting that the 3′ end of the EST is the real end of the transcript. The longest ORF starts at nt 11 and is 105 amino acids long. The 5′ UTR sequence is 10 nt, and the 3′ UTR is 271 nt and contains a poly(A) signal. A sequenced PCR clone showed a deletion of the last 23 nt of exon 2 (nt 291–313 inclusive), in correspondence with a GT dinucleotide: should this correspond to a real alternative transcript, the ORF of the last portion of the gene will be changed. However, this deletion was not present in any of the four cEST clones (Fig. 4). PCR amplification from YAC 780 DNA, which is known to contain the GdX gene (Palmieri *et al.*, 1994; see Fig. 2), showed that

ITBA1

```
   1.  gaatttcccccagcgaggcgagtgaggcgaaatacccgtatggtgatagctggccttttcgcgccaatactgaaaaaggcagaacgttcctccgctggcg
 101.  ccagccaatcagcaggactcctgcttccttcggggcaaggtcgcagcatctgcctcggaaatcacgaaatcacgggggcttctttctgctggctcagccg
 201.  ggaggcccagagtgttctgcagaggctgcgtattgaaggctgctctctgaagctccctgccccaggtcacgccgccggttccagATGAATCCAGAGTGGG
   1.                                                                                    M  N  P  E  W

 301.  GGCAGGCCTTCGTGCACGTGGCCGTGGCCGGTGGCCTCTGTGCCGTGGCTGTGTTCACGGGCATTTTCGACAGTGTTTCCGTGCAAGTGGGCTATGAGCA
   6.  G  Q  A  F  V  H  V  A  V  A  G  G  L  C  A  V  A  V  F  T  G  I  F  D  S  V  S  V  Q  V  G  Y  E  H

 401.  CTACGCCGAGGCGCCCGTGGCCGGCCTCCCTGCCTTCCTGGCCATGCCGTTCAACTCACTCGTGAACATGGCCTACACGCTGCTGGGGCTGTCGTGGCTG
  40.    Y  A  E  A  P  V  A  G  L  P  A  F  L  A  M  P  F  N  S  L  V  N  M  A  Y  T  L  L  G  L  S  W  L

 501.  CACAGGGGCGGCGCGATGGGGCTGGGTCCCCGCTACCTGAAGGACGTGTTCGGCAGGCCATGGCCCTGCTCTATGGCCCCGTGCAGTGGCTGCGCCTGTGGA
  73.   H  R  G  G  A  M  G  L  G  P  R  Y  L  K  D  V  F  A  A  M  A  L  L  Y  G  P  V  Q  W  L  R  L  W

 601.  CGCAGTGGCGCCGTGCCGCGGTGCTGGACCAGTGGCTCACACTGCCCATCTTTGCATGGCCCGTGGCCTGGTGCCTCTACCTAGACCGCGGCTGGCGGCC
 106.  T  Q  W  R  R  A  A  V  L  D  Q  W  L  T  L  P  I  F  A  W  P  V  A  W  C  L  Y  L  D  R  G  W  R  P

 701.  CTGGCTGTTCCTCTCTCTTGAGTGCGTCTCCCTGGCCAGTTATGGCCTCGCTCTGCTGCATCCCCAGGGCTTCGAGGTCGCACTGGGTGCTCACGTGGTG
 140.    W  L  F  L  S  L  E  C  V  S  L  A  S  Y  G  L  A  L  L  H  P  Q  G  F  E  V  A  L  G  A  H  V  V

 801.  GCCGCTGTGGGGCAGGCGCTGCGCACCCACAGGCACTATGGCAGCACCACCTCGGCTACCTACTTAGCTTTGGGGGTGCTCTCTTGCCTGGGCTTTGTGG
 173.   A  A  V  G  Q  A  L  R  T  H  R  H  Y  G  S  T  T  S  A  T  Y  L  A  L  G  V  L  S  C  L  G  F  V

 901.  TCCTCAAGCTGTGTGACCATCAGCTCGCACGGTGGCGTCTCTTCCAGTGCCTCACAGGCCACTTCTGGTCCAAGGTCTGTGACGTGCTCCAGTTCCACTT
 206.  V  L  K  L  C  D  H  Q  L  A  R  W  R  L  F  Q  C  L  T  G  H  F  W  S  K  V  C  D  V  L  Q  F  H  F

1001.  TGCGTTTTTGTTTCTGACGCATTTCAACACTCACCCAAGATTCCATCCCTCTGGCGGGAAGACGCGTTGAacccagggaagaacctgctgaaaaccgatg
 240.   A  F  L  F  L  T  H  F  N  T  H  P  R  F  H  P  S  G  G  K  T  R  *
1101.  acccccagcattgaaatggactctgagatggcagcgtggtgccagtgtcagacatcctgtgtgtgatgatatgcactgattacacaagactgccctttcc
1201.  tgagaagctgcgggcttcggtgtggaggggtggagtgctgtgatctcgacaacttactttcaaagacataaagcacagatctccgcacaggggatgtgtg
1301.  tgttcctgatgtaatttgcataacttttctgtagtttgaaatgtttccaaataaatattggcaaggggagtggaaatgacaccaagaagcccctcatgct
1401.  catggttggac
```

ITBA2

```
   1.  cgggacgcggATGCAGACGCAGGCGGAGGCGCTGACGGCGGGGATGGCCGGGGTGGCCACAGCTGCCGCGGGGGCGTGGACACAGCCGCAGCTCCGGCCG
   1.              M  Q  T  Q  A  E  A  L  T  A  G  M  A  G  V  A  T  A  A  A  G  A  W  T  Q  P  Q  L  R  P
                                                                                          ↓
 101.  GTGGAGCTCCCCCAGCGCACGCGCCAGGTCCGGGCAGAGACGCCGCGTCTGCCGCAGGGGGTCACGAATGCGGCCGCACATATTCACCCTCAGCGTGCCT
  31.   V  E  L  P  Q  R  T  R  Q  V  R  A  E  T  P  R  L  P  Q  G  V  T  N  A  A  A  H  I  H  P  Q  R  A

 201.  TTCCCGACCCCCTTGGAGGCGGAAATCGCCCATGGGTCCCTGGCACCAGATGCCGAGCCCCACCAAAGGGTGGTTGGGAAGGATCTCACAGTGAGTGGCA
  64.   F  P  D  P  L  G  G  G  N  R  P  W  V  P  G  T  R  C  R  A  P  P  K  G  G  W  E  G  S  H  S  E  W  Q
                ↓
 301.  GGATCCTGGTCGTCCGCTGGAAAAGCTGAagactgtcgcctgctccgaatttccgtcatcaactttcttgaccagctttccctggtggtgcggaccatgca
  98.    D  P  G  R  P  L  E  S  *
 401.  gcgctttgggccccccgtttcccgctaagcctggcctgggcaaatggagcgaggtcccactttgcgtctccttgtaggcagtgcgtccatccttccctag
 501.  ggcaggaattcccacagttgctacttttcctgggagggcctcatgttttatctggttcttaaatgtttgttactacagaaaataaaactgaggtattatt
```

FIG. 1. Sequences of ITBA1 (top) and ITBA2 (bottom) cDNA. The ATG start codon and the poly(A) signal are in boldface. Capital letters indicate the putative translated portions, while lowercase letters indicate the 5′- and 3′-untranslated regions. Both the sequences contained a poly(A) tail (not shown). The translation of the longest ORF of the two genes is shown in the one-letter amino acid code. The arrows indicate the splicing sites.

ITBA2 is contained in a very limited genomic region of about 1150 bp. By comparing the cDNA and genomic sequences, three exons of 184, 129, and 286 bp were identified, with two small introns, whose boundaries follow the GT/AG rule. Therefore the EST clone represents a true spliced transcript. Obvious homologies were not detected in databases.

The sequence data are in keeping with the transcript size as detected with the Northern blot, although it must be pointed out that the exact begin-

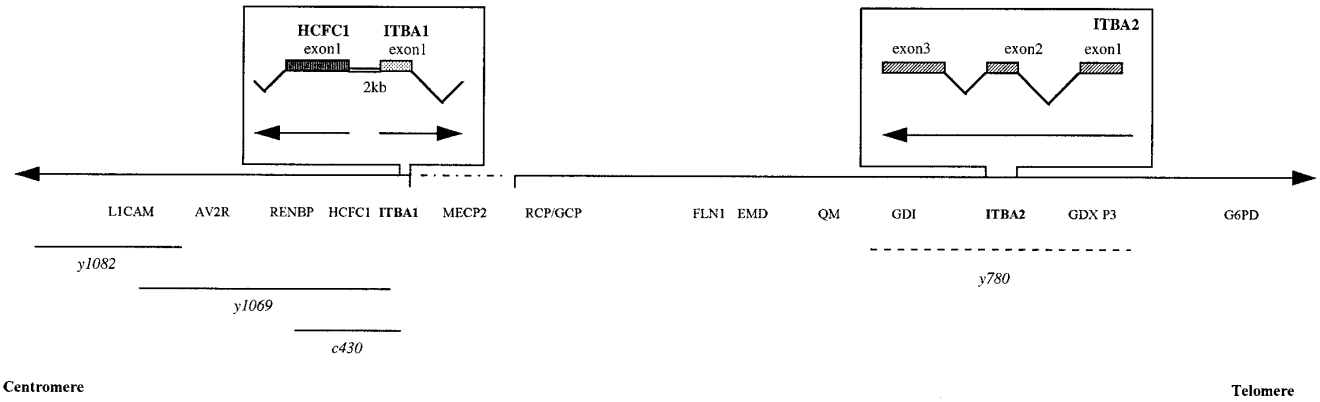

FIG. 2. Fine mapping of ITBA1 and ITBA2 in Xq28. The solid line indicates the genomic region contained between the L1CAM and G6PD genes; the discontinuous region is indicated by the gap between the two contigs, which have not yet merged (Nelson *et al.*, 1995). Some of the genes already identified in this region are listed below the solid line, as reference markers (not in scale). Numbers below indicate YAC (y) or cosmid (c) as described in Frattini *et al.* (1993). The dotted y780 is a YAC (containing a genomic deletion) from which ITBA2 genomic sequence was amplified. Boxes show the direction of the transcription of both ITBA1 and ITBA2 as well as the distance between HCFC1 and ITBA1 (left) and the genomic structure of ITBA2, which lies approximately 5 kb from the end of GdX (the latter is transcribed from telomere to centromere).

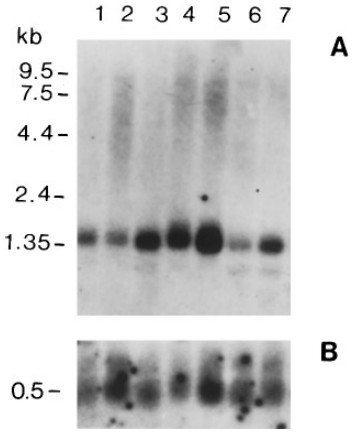

FIG. 3. Ubiquitous expression of ITBA1 (A) and ITBA2 (B). Marker sizes are indicated to the left. Lane 1: spleen; lane 2: thymus; lane 3: prostate; lane 4: testis; lane 5: ovary; lane 6: small intestine; lane 7: colon.

ning of the transcript has not been established. Like ITBA1, ITBA2 appears to be ubiquitously expressed (Fig. 3B), but does not seem to be extremely conserved in evolution, especially compared to other genes from the same region (Frattini *et al.*, 1995). A signal was detected in great apes, cows, and horses, but not in mice.

DISCUSSION

Here we report the characterization of two new genes in Xq28, the first in the L1CAM region and the second in the G6PD region. Both these genomic segments are among the best-characterized portions of the human genome at the transcriptional level, but although they have been extensively investigated by various groups, these genes have not been described in the various transcription unit maps that have been published so far (Tribioli *et al.*, 1994; Bione *et al.*, 1993; Sedlacek *et al.*, 1993; Palmieri *et al.*, 1994). The existence of a gap between the L1CAM and the CV contigs, which has only recently been reduced in size but is not yet closed (Nelson *et al.*, 1995), could explain the difficulty in cloning ITBA1, while the small size of the ITBA2 transcript and the fact that this region is very abundant in genes could have led researchers to think that this region could not contain other genes in addition to those already identified.

Both ITBA1 and ITBA2 were found with an alternative approach based on sequencing of genomic clones that have a high probability of representing CpG islands, followed by a computer search of the new EST databases. We think that, in the near future, this strategy will be implemented on a large scale, allowing easy mapping of many genes. With an approach similar to

```
yu55r1   >                                                                                                  CGGGACGNGANTGCAGACGCAGGCGNAGG.GCTGACGGCGGGGAT
ITBA2                                                                                                                                    GGAGGCGCTGACGGCGGGGAT
genITBA2 < AAGGAGCCACAGAGAGGGCGGGGCGTAGGACCTGCGCTTCGGGGGTGGAGTCGGAGCGGCGCGGCGGCGGTCATGCGGGACGCGGATGCAGACGCAGGCGGAGGCGCTGACGGCGGGGAT

yi95r1   <                                                                                                                                                                AAAGGAAA
ye56r1   >  GCNGAGGGTGGCATCAGCTGCCGCGGNGGGCGTGGACACAGCCGCAGCTCCGGCTCGGTGGAGCTCCCCCAGCGCACG.GCAGAGTCCGGGCAGAGACGCCGCGTCT.C.NANGNGGGG
yu55r1   > GGCCG.GGGTGGC.ACAGCTGCCGCGG.GGG.NTGGACACAGCCGCAGCTCCGGC.CGGTGGAGCTCCCCCAGCGCACG.GNCAGGTCCGGGCAGAGACGCCGCGTCTGC..CANGGGGG
ITBA2      GGCCG.GGGTGGCCACAGCTGCCGCGG.GGGCGTGGACACAGCCGCAGCTCCGGC.CGGTGGAGCTCCCCCAGCGCACGCGCCAGGTCCGGGCAGAGACGCCGCGTCTGCGGCCAGGGGG
genITBA2 < GGCCG.GGGTGGCCACAGCTGCCGCGG.GGGCGTGGACACAGCCGCAGCTCCGGC.CGGTGGAGCTCCCCCAGCGCACGCGCCAGGTCCGGGCAGAGACGCCGCGTCTGCGGCCAGGGGG

ye56s1   <                                                                                                                                                                CACCNA
yu55s1   <                        AAA.....ATNACCCTNAACGTG.CCTTTCCCGAACCCCCTTGGAAGCGGAAAAATCG.CCAAGGGTCCCCTGGCAACCAGATGCCG.AAGCCCCACC..
yi95s1   >                       CAT.A.....TTCACCCTCAGCGTG.CCTTTCCCG.ACTNCCTTTGAGGCGG..AAATCG.CCATGGGT.CCCTGGC.ACCAGATGCCG..AGCCCCACC..
yt85s1   >          TGGTACCG.CACATNA.....TTCACCCTCAGCGTG.CCTTT.NCNAANANCCTTAGAGGTGG.GAAATTGCCCATGGGTCCCTGGGC.TCCAGATGCCG..AACCCCACC..
yi95r1   < TAA..ACTTCGGCCG.CAACA.A.....ATCACCCTNAACGTGCCCTTTCCCGAACCCCCTTGGAANCGG.AAAATCGCCCANGGGTCCCCTGGC.ACCAAATGCCGAAACCCCNACC..
ye56r1   > TCACGAATGCGGCCGACATATTACCCCCAGCACCCTCAGCGTG.CCTTTCCCG.ACCCCCTTGGAGGCGG..AAATCG.CCATGGGTCCCTGGGC.ACCAGATGCCG..AGCCCCACC..
yu55r1   > TCACGAATGCGGCCG.CACAT.A.....TTCACCCTCAGCGTG.CCTTTCCCG.ACCCCCTTGGAGGCGG..AAATCGCCCATGGGT.CCCTGGC.ACCAGATGCCG..AGCCCCACC..
ITBA2      TCACGAATGCGGCCG.CACAT.A.....TTCACCCTCAGCGTG.CCTTTCCCG.ACCCCCTTGGAGGCGG..AAATCGCCCATGGGT.CCCTGGC.ACCAGATGCCG..AGCCCCACC..
genITBA2 < TCACGAATGCGGCCG

ye56s1   < AAAGGGTGGTTGGGAAAAGGAATCTNAACAAGTGAAGTGGNAAGGATCCCTGGTCCGTCCCGCTGGAAAAAGCTGAAAGACTGTCGCCTGNTCCGAATTTCCGTCATCAACTTTCTTGAC
yu55s1   < AAAGGGTGGTTGGG..AAGG.ATCT.CAC.AGTGAAGTGGC.AGGANCCT.GGT.CGT.CCGCTGG..AAAGCTG.AAGACTGTCGCCTGCTCCGAATTTCCGTCATCAACTTTCTTGAC
yi95s1   > AAAGGGTGTTTGGG..AAGG.ATCTG.AC.AGTG.AGTGGC.AGGATCCTGGGT.CGT.CCGCTGG..AAAGCTG.AAGACTGTCGCCTGCTCCGAATTTCCGTCATCAACTTTCTTGAC
yt85s1   > AAAGGGTGGTTGGG..AAGG.ATCT.CAC.AGTG.AGCGGC.AGGATCCTGGGC.CGT.CCGCTGG..AAAGCTG.AAGACTGTCGCCTGCTCCGAATTTCCCTTATCAACTTTCTCGAC
yi95r1   < AAAGGGTGGTTGGG..AAGG.ATCTGAAC.AGTG.AGTGGC.AGGATCCT.GGT.CGT.CCGCTGG..AAAGCTG.AAGACTGTCGCCTGCTCCGAATTTCCGTCATCAACTTTCTTGAC
ye56r1   > AAAGGGTGGTTTGGGGAAGG.ATCT.CAC.AGTTGAGTGGGCAGGATCCTGGGT.CGT.CCGCTGG..AAAGTTGAAAGATTG
yu55r1   > AAAGGGTGGTTNGG..AAGG.ATCT.CAC.AGT
ITBA2      AAAGGGTGGTTGGG..AAGG.ATCT.CAC.AGTG.AGTGGC.AGGATCCT.GGT.CGT.CCGCTGG..AAAGCTG.AAGACTGTCGCCTGCTCCGAATTTCCGTCATCAACTTTCTTGAC

ye56s1   < CAGCTTTCCCCTGGT.GGTGCGGACCATGCAGCGCTTTGGGCCCCCCGTTTCCCGCTAAGCCT.GGCCT.GGGCAAAT.GGAGCGAGGTCCCACTTTGCGTCTCCTTGTN.GGCAGTGCG
yu55s1   < CAGCTTTCCCT.GGT.GGTGCGGACCATGCAGCGCTTTGGGCCCCCCGTTTCCCGCTAAGCCT.GGCCT.GGGCAAAT.GGAGCGAGGTCCCACTTTGCGTCTCCTTGTA.GGCAGTGCG
yi95s1   > CAGCTTTCCCTGGGTGGGTGCGGACCATGCAGCGCTTTGGGCCCCCCGTTTCCCGCTAAGCCTGGGGCCTGGGGCAAATGGGAGCGAGGTCCCACTTTGCGTCTCCTTGTAGGGCAGTGCG
yt85s1   > CATCTTTCTCT.GGT.NGTGCGGACCCTGCAGCACTTTNGGCCCCCAGTTTCCCGCTAAGCCT.NGCCT.NGGAAAAT.GGGGCGAAGTCTAACCTTGCGTCTTCTCCTA.GGCAGTGCA
yi95r1   < CAGCTTTCCCT.GGT.GGTGCGGACCATGCAGCGCTTTGGGCCCCCCGTTTCCCGCTAAGCCT.GGCCT.GGGCAAAT.GGAGCGAGGTCCCACTTTGCGTCTCCTTGTA.GGCAGTGCG
ITBA2      CAGCTTTCCCT.GGT.GGTGCGGACCATGCAGCGCTTTGGGCCCCCCGTTTCCCGCTAAGCCT.GGCCT.GGGCAAAT.GGAGCGAGGTCCCACTTTGCGTCTCCTTGTA.GGCAGTGCG

ye56s1   < TCCATCCTTCCCTA.GGGCAGG.AATTCCCACAG.TTGCT.ACTTTCCT.GGGAGGGCCTCATGTTTT..ATCTGGTTCTT..AAATG.TTTGT.TACT.ACAGAAAATAAA
yu55s1   < TCCATCCTTCCCTA.GGGCAGG.AATTCCCACAG.TTGCT.ACTTTCCT.GGGAGGGCCTCATGTTTT..ATCTGGTTCTT..AAATG.TTTGT.TACT.ACAGAAAATAAAACTGCGCT
yi95s1   > TCCATCCTTCCCTAGGGGCAGGAATTTCCCACAGTTTGNTAATTTTCCTGGGGAGGGNCTCATGTTTTTAATCTNGTTCTTTAAAATGTTTTGT.TA
yt85s1   > TCCATCC.TNCCTA.GGGCAGGTAATTTCCACAG.TTGCT.ACTTTCCC.GGGA.GGCCTTATGGTCG..GCCT.GTACT.....GCC.TTTGTGTGCTCAGGGGTNCCTTGTTTTCTCT
yi95r1   < TCCATCCTTCCCTA.GGGCAGG.AATTCCCACAG.TTGCT.ACTTTCCT.GGGAGGGCCTCATGTTTT..ATCTGGTTCTT..AAATG.TTTGT.TACT.ACAGAAAATAAAACTGCGCT
ITBA2      TCCATCCTTCCCTA.GGGCAGG.AATTCCCACAG.TTGCT.ACTTTCCT.GGGAGG

yu55s1   < ACTACA
yt85s1   > NGTTCTTAAATGTTTNTTACCACAGAAAATAAAACTGAGGTATTATTAAAAAAAAAAAAAAAAAACCTCGTGCCCAATTTTTGGGAT
yi95r1   < ACT
```

FIG. 4. Assembling of the ITBA2-related sequences from the EST database (updated to November 1995). The final ITBA2 sequence is shown in boldface. GenITBA2 refers to the CpG island sequence used to probe the EST database. The "r1'" in the EST clone codes indicates sequences performed from the 5′ end, while the "s1'" indicates sequences obtained from the 3′ end of the cDNA clones from Soares' libraries.

the one suggested previously by us and others (Tribioli *et al.*, 1992), it should be possible to derive human genomic clones from rodent–human hybrids containing selected portions of the human genome by screening *Eag*I (or other rare-cutters) site-containing clones with radioactive total human DNA. Single-pass sequencing of these clones and comparison to EST databases should automatically localize all the detected EST identities to those specific genome portions. In addition, this would allow the isolation of genomic regions containing the promoter regions of these genes, since CpG islands are often located in the 5′ region of housekeeping genes (Cross *et al.*, 1994). The localization of the ITBA1 and ITBA2 genes is a clear example of the potentiality of this approach, and we have further evidence of the localization of other genes to the Xq24–qter region using this method (A. Frattini, unpublished observation). However, it must be pointed out that this approach is dependent on the quality of the cDNA libraries used to generate ESTs and on the way they are constructed, since CpG islands are found prevalently (but not exclusively) at the 5′ ends of the genes, which should be present in the EST libraries to maximize the applicability of this approach.

As expected from their association with CpG islands, both genes are ubiquitously expressed. ITBA1 and ITBA2 do not seem very conserved in evolution, as judged from zoo blot analysis and from preliminary screening and PCR amplification attempts in mice. The functions of both genes are at present unknown, but their small size and the absence of homologies, as exemplified by the Emerin gene (Bione *et al.*, 1994), do not exclude *a priori* their possible involvement in one of the diseases mapped to the L1CAM/G6PD interval whose responsible gene has not been identified so far.

ACKNOWLEDGMENTS

We are grateful to Professor R. Dulbecco for his suggestions and careful reading of the manuscript and to Professor L. Rossi Bernardi for his encouragement. The technical assistance of Dario Strina and Lucia Susani is also gratefully acknowledged. We also thank Sophie Bevan for her careful typing of the manuscript. Clones HSCZXE10 and HSBB5D05 were obtained from the Genexpress cDNA Program, Laboratoire Généthon, Evry, France. This work was supported by grants from CNR (P. F. Ingegneria Genetica to P.V.), from Telethon Italy (to I.Z.), and from Biomed Program, E. C. no. BIOG-CT95-0226 (to L. M.).

REFERENCES

Altschul,S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.* 215, 403–410.

Bione, S., Tamanini, F., Maestrini, E., Tribioli, C., Poutska, A., Torri, G., Rivella, S., and Toniolo, D. (1993). Transcriptional organization of a 450 kb region of the human chromosome in Xq28. *Proc. Natl. Acad. Sci. USA* 90, 10977–10981.

Bione, S., Mestrini, E., Rivella, S., Mancini, M., Regis, S., Romeo, G., and Toniolo, D. (1994). Identification of a novel X-linked gene responsible for Emery–Dreifuss muscular dystrophy. *Nature Genet.* 8, 323–329.

Cross, S. H., Charlton, J. A., Nan, X.,. and Bird, A. P. (1994). Purification of CpG islands using a methylated DNA binding column. *Nature Genet.* 6, 227–235.

Faranda, S., Frattini, A., and Vezzoni, P. (1995). The human genes encoding renin-binding protein and host cell factor are closely linked in Xq28 and transcribed in the same direction. *Gene* 155, 237–239.

Feinberg, A., and Vogelstein, B. (1983). A technique for radiolabeling DNA restriction endonuclease fragments to high specific activity. *Anal. Biochem.* 132, 6–13.

Frattini, A., Faranda, S., Redolfi, E., Zucchi, I., Villa, A., Patrosso, M. C., Strina, D., Susani, L., and Vezzoni, P. (1994). Genomic organization of the human VP16 accessory protein, a housekeeping gene (HCFC1) mapping to Xq28. *Genomics* 23, 30–35.

Frattini, A., Zucchi, I., Villa, A., Patrosso, M. C., Repetto, M., Susani, L., Strina, D., Redolfi, E., Vezzoni, P., Romano, G., Palmieri, G., Esposito, T., and D'Urso, M. (1993). Type 2 vasopressin receptor gene, the gene responsible for nephrogenic diabetes insipidis, maps to Xq28 close to the L1CAM gene. *Biochem. Biophys. Res. Commun.* 193, 864–871.

Nelson, D., Ballabio, A., Cremers, F., Monaco, A. P., and Schlessinger, D. (1995). Report of the sixth international workshop on X chromosome mapping. *Cytogenet. Cell. Genet.*, in press.

Palmieri, G., Romano, G., Ciccodicola, A., Casamassimi, A., Campanile, C., Esposito, T., Capra, V., Lania, A., Johnson, S., Reinbold, R., Poutska, A., Schlessinger, D. and D'urso, M. (1994). YAC contig organization and CpG island analysis in Xq28. *Genomics* 24, 149–158.

Patrosso, M. C., Repetto, M., Villa, A., Frattin, A., Faranda, S., Mancini, M., Maestrini, E., Toniolo, D., and Vezzoni, P. (1994). The exon–intron organization of the human X-linked gene (FLN1) encoding actin-binding protein 280. *Genomics* 21, 71–76.

Pilia, G., Little, D. R., Aissani, B., Bernardi, G., and Schlessinger, D. (1993). Isochores and CpG islands in YAC contigs in Xq26.1–qter. *Genomics* 17, 456–462.

Sedlacek, Z., Korn, B., Konechi, D. S., Siebenhaar, R., Coy, J. F., Kioschis, P., and Poustka, A. (1993). Construction of a transcription map of a 300 kb region around the human G6PD locus by direct cDNA selection. *Hum. Mol. Genet.* 2, 1865–1869.

Tribioli, C., Tamanini, F., Patrosso, M. C., Milanesi, L., Villa, A., Pergolizzi, R., Maestrini, E., Rivella, S., Bione, S., Mancini, M., Vezzoni, P., and Toniolo, D. (1992). Methylation and sequence analysis around EagI sites: identification of 28 new CpG islands in Xq24–q28. *Nucleic. Acids. Res.* 20, 727–733.

Tribioli, C., Mancini, M., Plassart, E., Bione, S., Rivella, S., Sala, C., Torri, G., and Toniolo, D. (1994). Isolation of new genes in distal Xq28: Transcriptional map and identification of a human homolog of the ARD1 *N*-acetyl transferase of Saccharomyces cerevisiae. *Hum. Mol. Genet.* 3, 1061–1067.

Wilson, A. C., Parrish, J. E., Massa, H. F., Nelson, D. L., Trask, B. J., and Herr, W. (1995). The gene encoding the VP16-accessory protein HCF (HCFC1) resides in the human Xq28 and is highly expressed in fetal tissues and the adult kidney. *Genomics* 25, 462–468.

Zoppè, M., Frattini, A., Faranda, S., and Vezzoni, P. (1996). The complete sequence of the Host cell factor 1 (HCFC1) gene and its promoter: A role for YY1 transcription factor in the regulation of its expression. *Genomics*, in press.